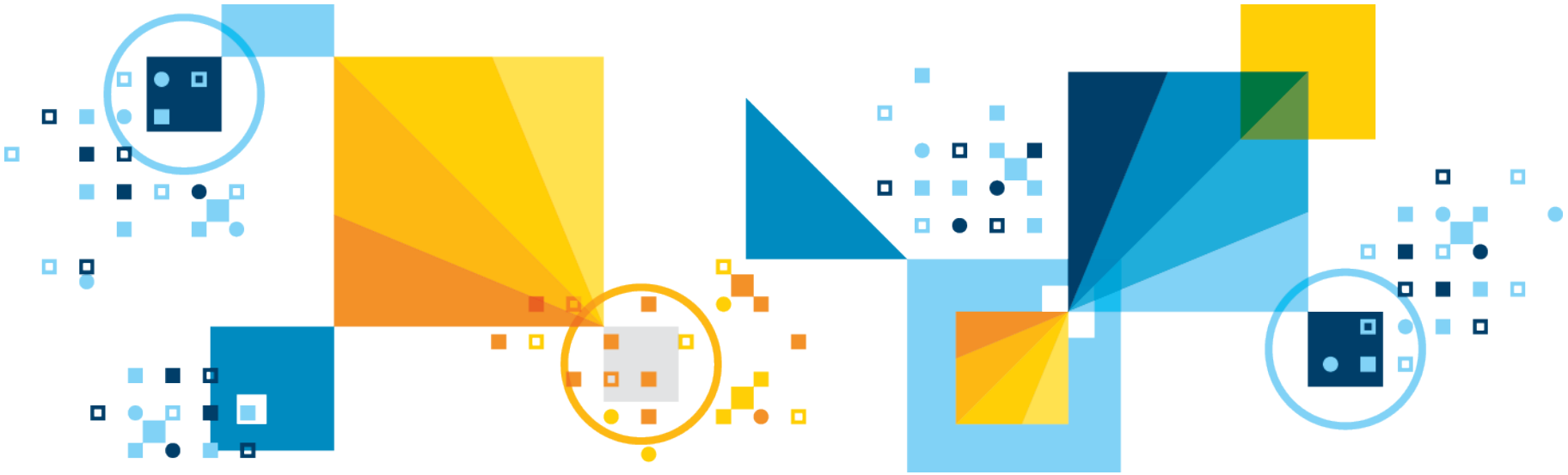
Dec 15th, 2016

Welcome to the IBM IIS Tech Talk

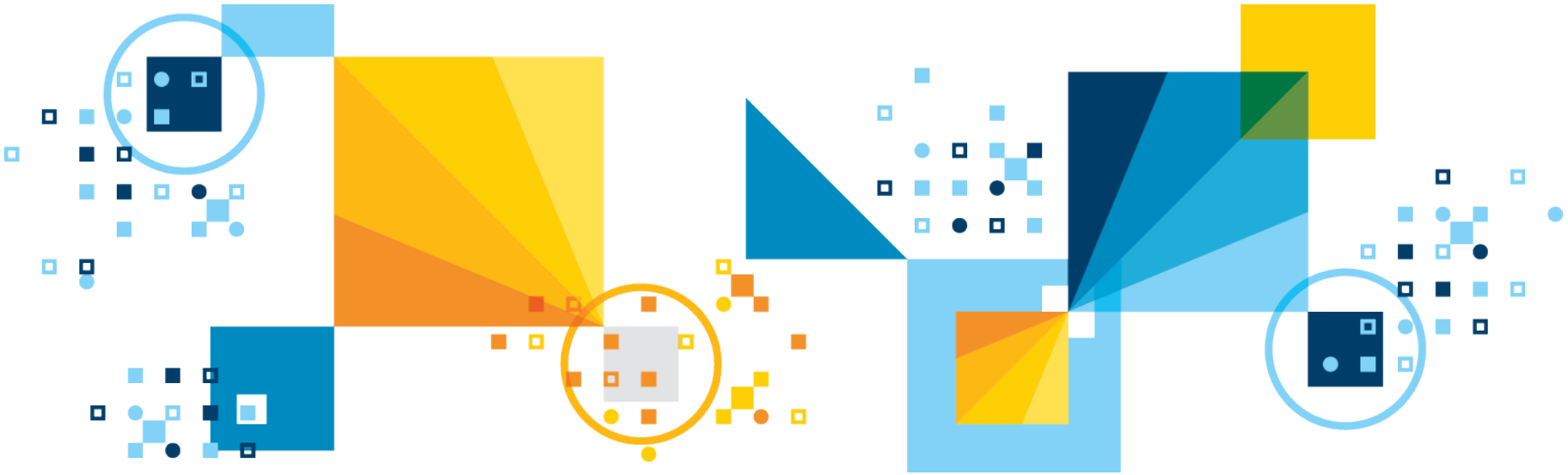
Data Quality in Information Analyzer



Dec 15th, 2016

Information Analyzer Data Quality Deep Dive

Yannick Saillet – Software Architect



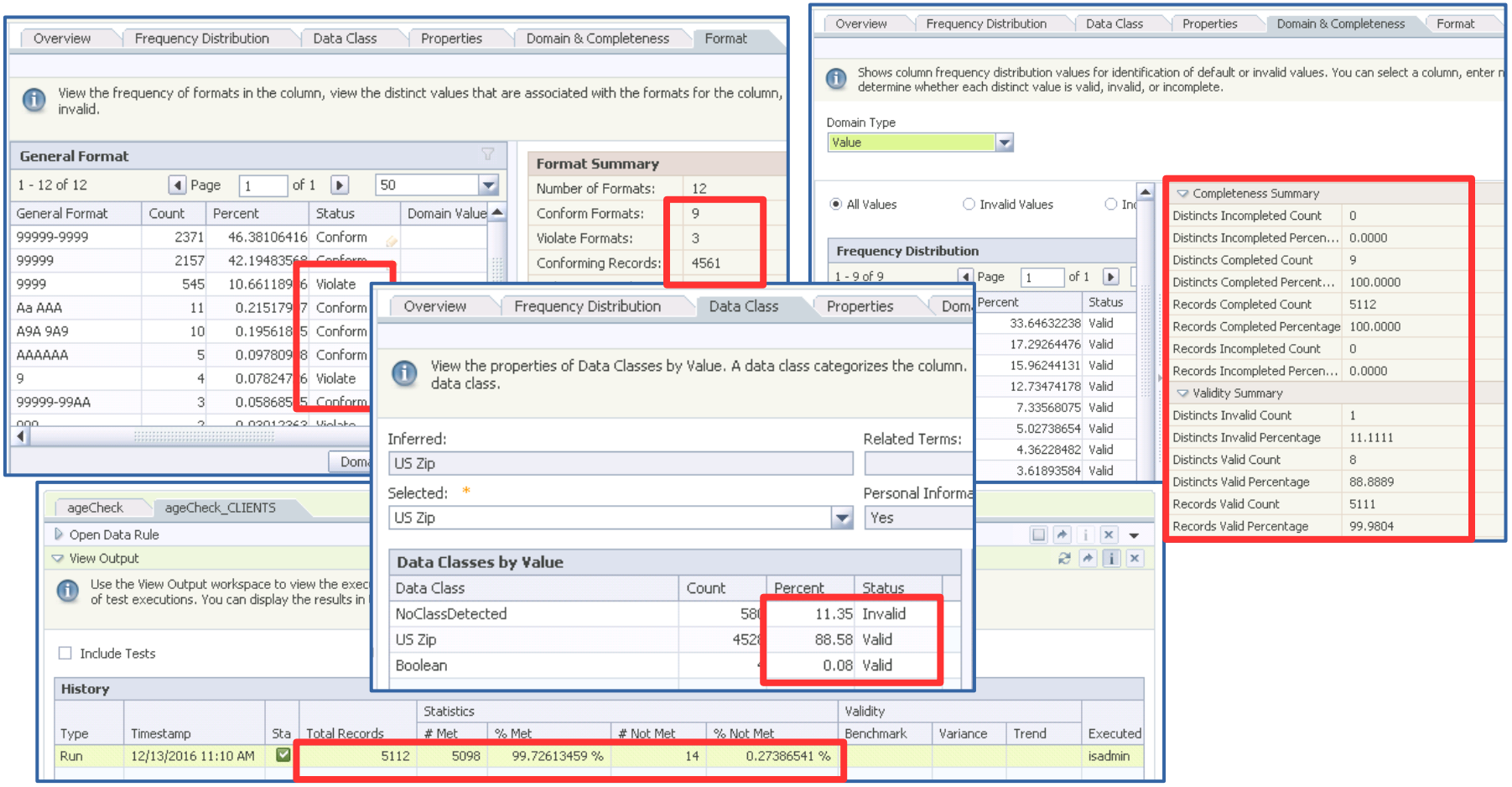
AGENDA

- Data Quality in Information Analyzer in the past
- Data Quality in IA 11.5 thin client
- The Data Quality Framework
- Out of the box data quality problems detection
- Reporting
- Q&A

Analytics Platform Services - Disclaimer

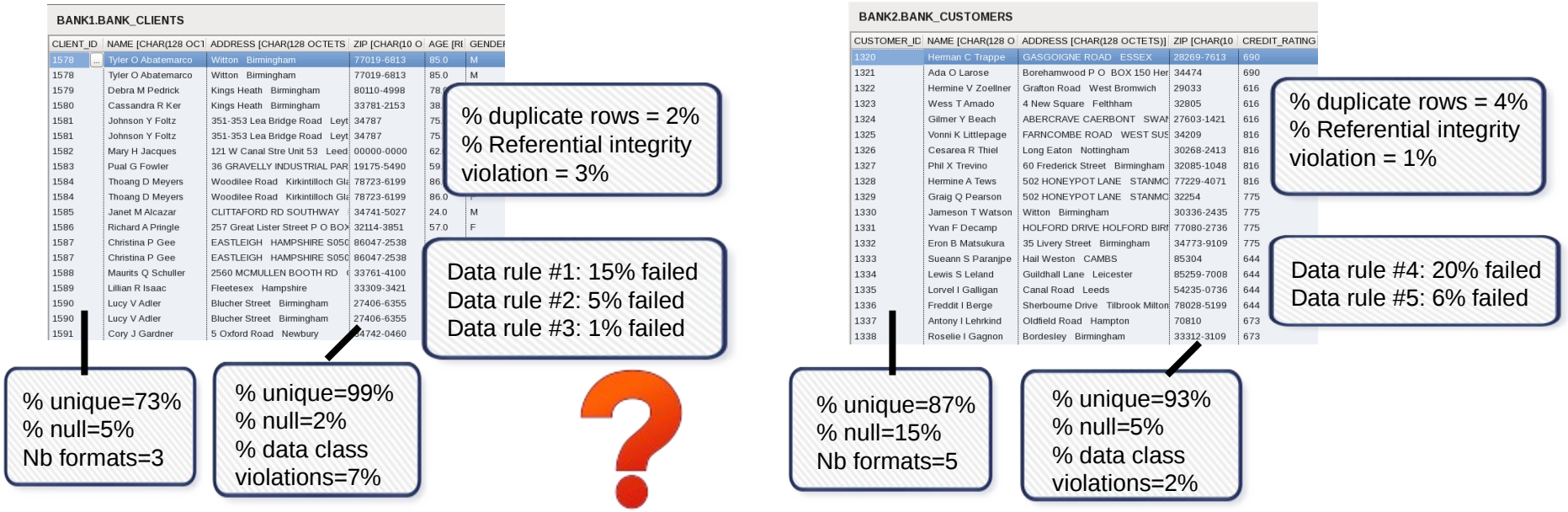
- IBM’s statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM’s sole discretion.
- Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.
- The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.
- The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.
- Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user’s job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.

Data Quality in Information Analyzer in the past



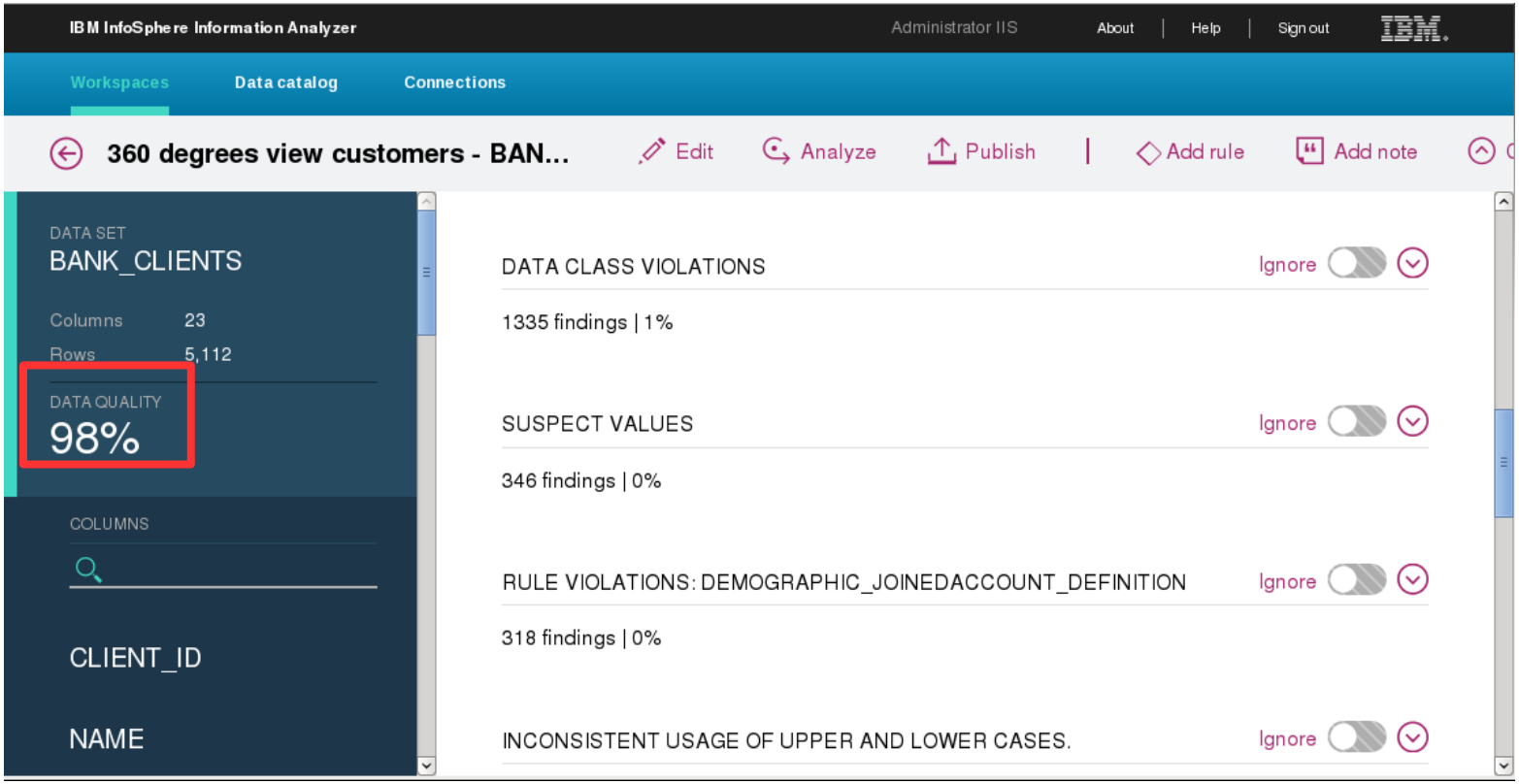
- Many data quality related information widespread on many different screens:
  (Domain validity, formats, data classes, data rules, key referential integrity, ...)

The challenge of measuring Data Quality



- Which data set has the better quality?
- What are the quality problems I need to resolve first to increase the quality?
- => How to consolidate heterogeneous quality metric into a single metric?

Data Quality in Information Analyzer 11.5 Thin Client



- A **consolidated** single **data quality score** is computed for each data set and field.
- It makes it easy to compare the quality of different data sets or data fields.
- The full data quality details explaining the scores are still captured.

How it works: Data Quality Framework

- Extensible(*) framework where different algorithms specialized in detecting specific types of data quality problems can be plugged.
- The framework will run the algorithms on the data and consolidate the findings into:
  - A normalized data quality score for the whole data set, each data field, or even each row and value.
  - A distribution of the frequency of each type of problem is found for the whole data set or each individual data field.
- The data quality scores are computed in a standardized way
  - No matter how many algorithms are used.
  - No matter what type of problems these algorithms search.
  - No matter how the algorithms find those problems.

(*): As of the current version, the extension point to integrate new algorithms is not yet public. A public extension point may be considered in the future.

How it works: Data Quality Framework (cont)

- Each algorithm is a so called “Quality Scanner” which can search for the existence of one or multiple types of data quality problems within a single cell, a complete row, a complete column or the data set as a whole.
- The scanner declares to the framework:
  - The type of problems it searches,
  - Whether it search for problems within individual cells, rows, columns or the data set as a whole.
  - How many passes over the data it requires.
- The framework invokes the scanners.
- When a scanner detects a data quality problem, it annotates the cell, row, column or data set where it found the problem with:
  - The type of problem which has been identified.
  - The confidence that the problem is not a false positive.
- The framework consolidates the annotations produced by each scanner into consistent data quality scores.

Computing Data Quality Scores

- Data quality score
  - **Estimate of the proportion of reliable data values in the given dataset**
  - It is represented on the percentage scale

- Cell score: estimate proportion of all data quality problems per cell
  - *q*: data quality problem that is applicable for a given cell
  - *prev(q)*: prevalence for the data quality problem identified by q (prevalence=estimation of the % of values having that problem)
  - *conf(q)*: confidence for the data quality problem identified by q (confidence=probability that the detected problem is not a false positive)
  - Q_{cell}: the number of all data quality problems that are applicable for the given cell
  - *Score (cell):* data quality score computed for the given cell

$$Score(cell) = \prod_{q=0}^{Q_{cell}} (1 - prev(q) * conf(q))$$

Column and Total Data Quality Scores

- Column score: estimate the column proportion of reliable data values
  - *col*: data set column of interest
  - *row*: data set row numbered from 1 to *N*
  - *Score (col):* data quality score computed for the given column

$$Score(col) = \frac{1}{N} \sum_{row=1}^{N} Score(row, col)$$

- Dataset score: estimate of the total proportion of reliable data values
  - *M*: the number of columns in the given data-set
  - *weight(col)*: pre-specified weight for each column with the total *W*
  - Score (data-set): data quality score computed for the whole data-set

$$Score(dataset) = \frac{1}{W} \sum_{col=1}^{M} Score(col) * weight(col)$$

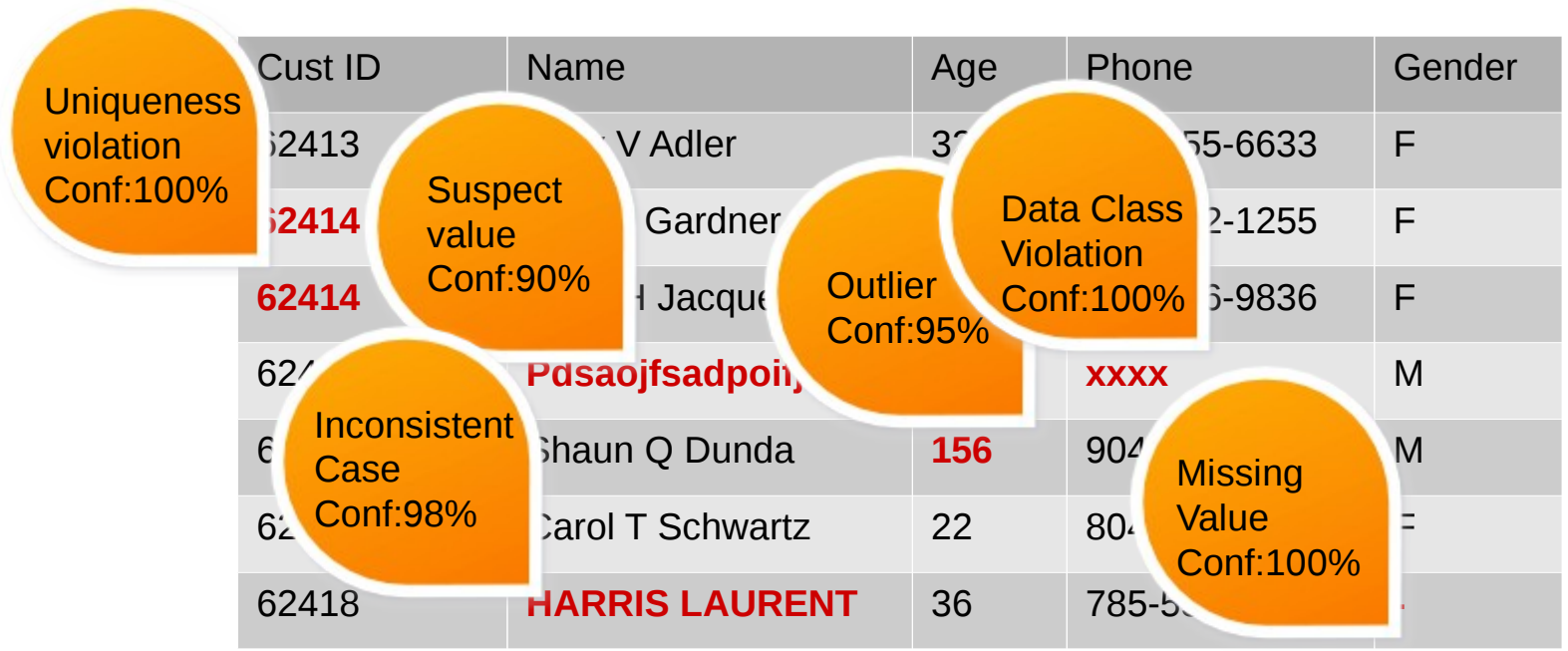
Example

Cust ID	Name	Age	Phone	Gender
62413	Lucy V Adler	32	334-555-6633	F
62414	Cory J Gardner	25	903-222-1255	F
62414	Mary H Jacques	18	777-156-9836	F
62415	Pdsaojfsadpoifj	46	xxxx	M
62416	Shaun Q Dunda	156	904-555-2940	M
62417	Carol T Schwartz	22	804-555-3164	F
62418	HARRIS LAURENT	36	785-555-5835	-

Example

Cust ID	Name	Age	Phone	Gender
62413	Lucy V Adler	32	334-555-6633	F
62414	Cory J Gardner	25	903-222-1255	F
62414	Mary H Jacques	18	777-156-9836	F
62415	**Pdsaojfsadpoifj**	46	**xxxx**	M
62416	Shaun Q Dunda	**156**	904-555-2940	M
62417	Carol T Schwartz	22	804-555-3164	F
62418	**HARRIS LAURENT**	36	785-555-5835	**-**

Example



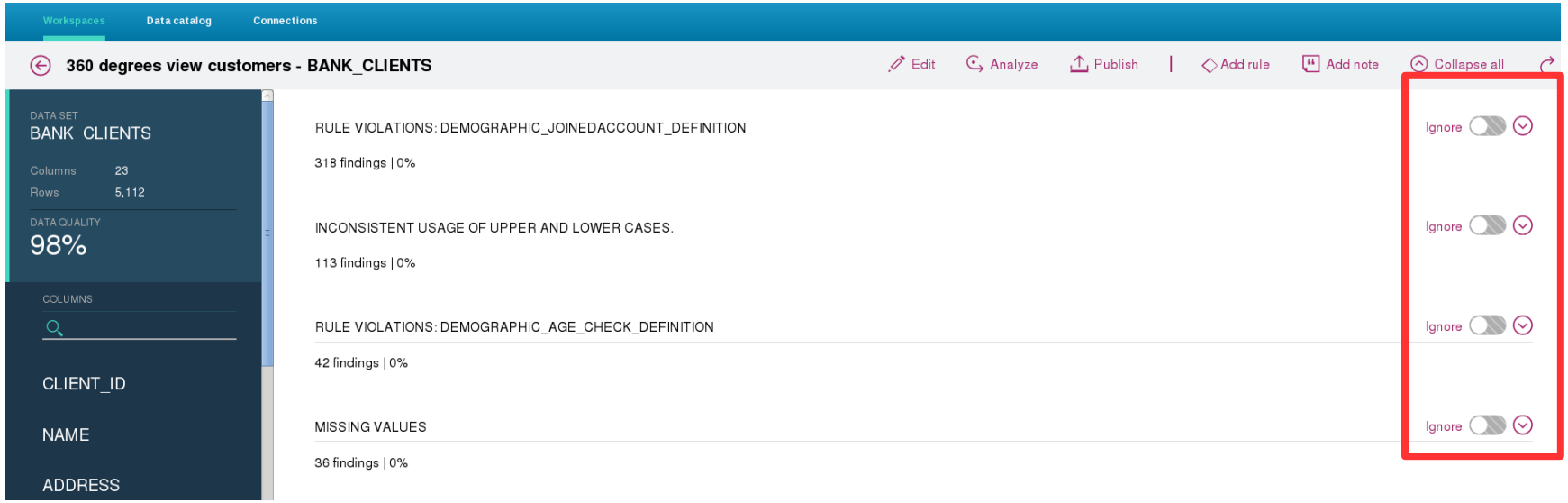
Example

Data Set Score: 80%

	Score: 71%	Score: 73%	Score: 86%	Score: 85%	Score: 85%
	Cust ID	Name	Age	Phone	Gender
	62413	Lucy V Adler	32	334-555-6633	F
	62414	Cory J Gardner	25	903-222-1255	F
	62414	Mary H Jacques	18	777-156-9836	F
	62415	**Pdsaojfsadpoifj**	46	**xxxx**	M
	62416	Shaun Q Dunda	**156**	904-555-2940	M
	62417	Carol T Schwartz	22	804-555-3164	F
	62418	**HARRIS LAURENT**	36	785-555-5835	**-**

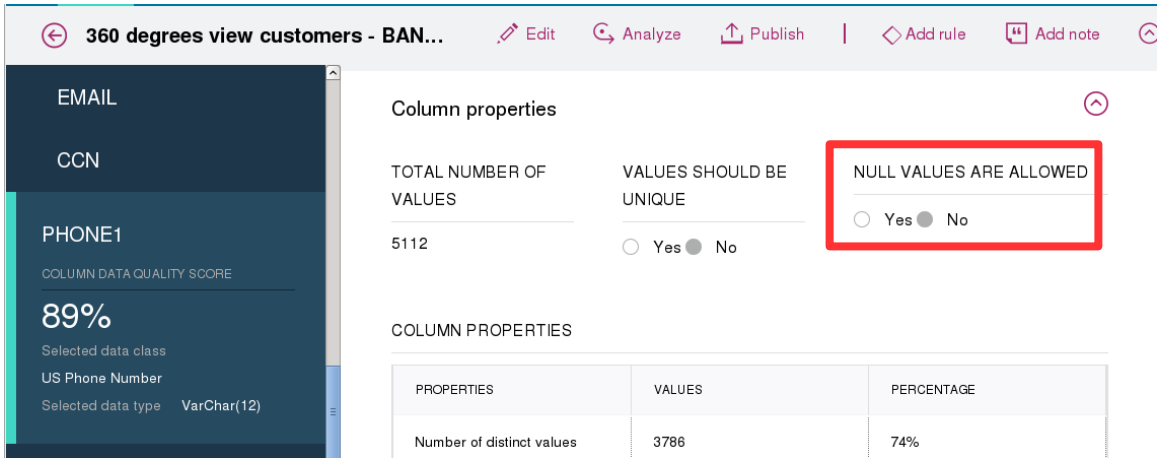
Data quality problems detected out of the box

- Information Analyzer includes a collection of out of the box algorithms specialized in detecting different types of data quality problems.
- You can choose to ignore some of these data quality problems in the UI.
  - Problems can be ignored for a single data field only, or for the whole data set.
  - Ignored problems won't impact the score and are not going to be searched in consequent analysis.



Data quality problems detected out of the box: Missing Values

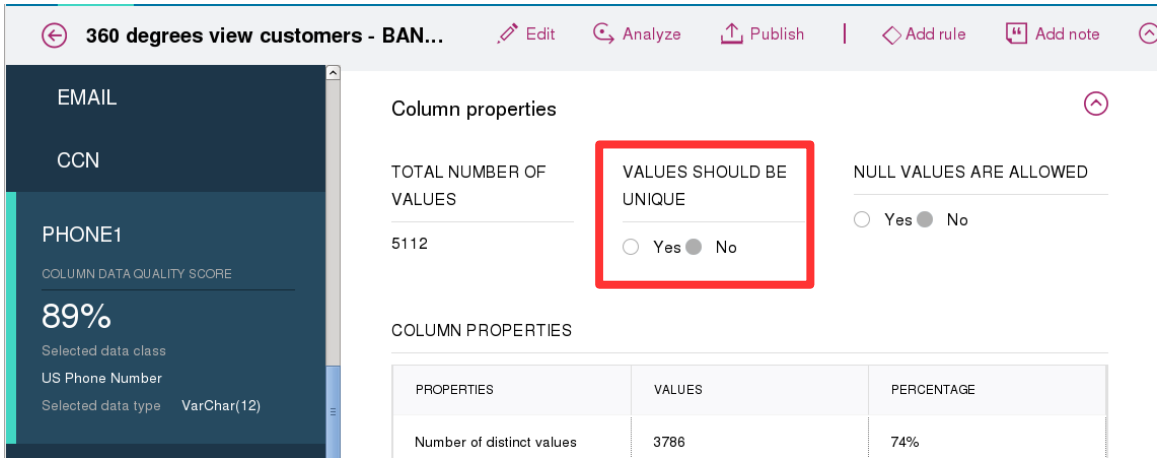
- Checks for missing values in data fields where missing values are not supposed to be expected.
- If a data field is declared by the user as non nullable, all null and empty values will be reported as a quality problem with a confidence of 100%.
- The nullability flag of the data fields can be controlled in the UI:



- If no information is available about the nullability of a data field, null and empty values are only reported as a quality problem if the majority of the values in the data field are not missing. The confidence of the reported problem will be proportional to the percentage of non missing values in the data field.

Data quality problems detected out of the box: Uniqueness Violations

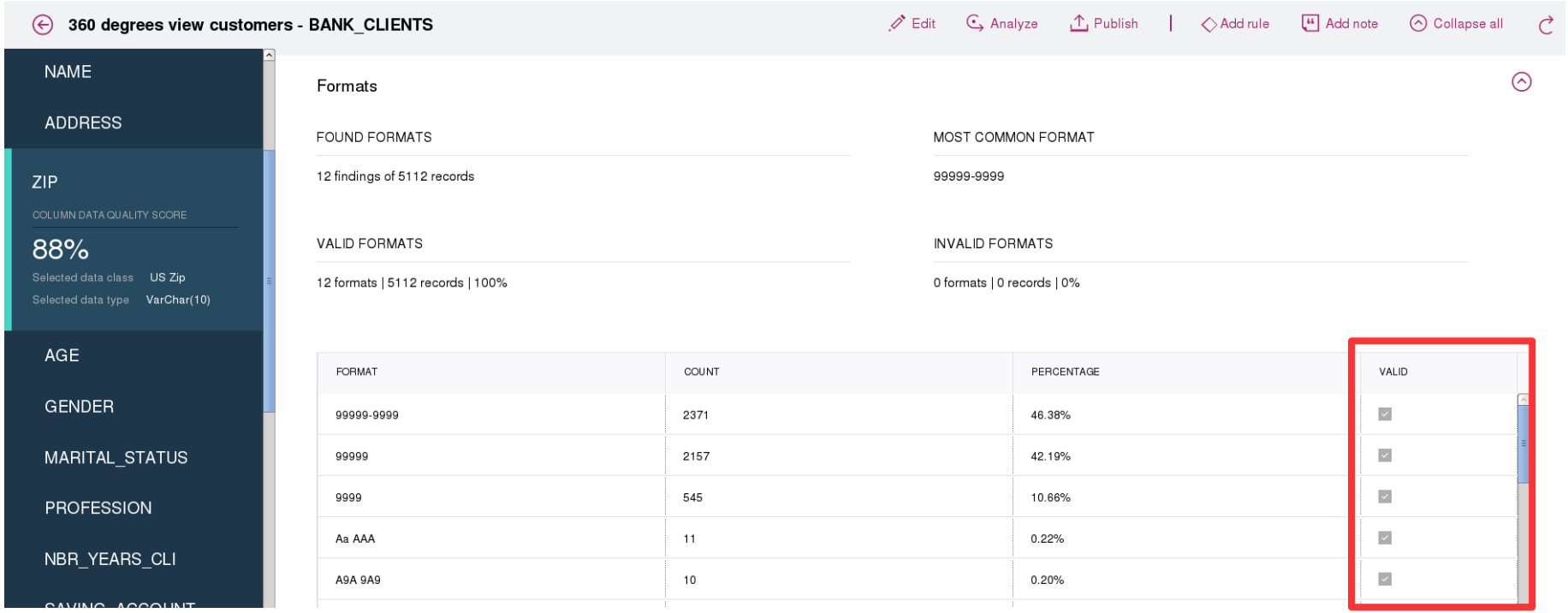
- Checks for duplicate values in data fields supposed to contain only unique values.
- If a data field is declared by the user as unique, all non unique values will be reported as a quality problem with a confidence of 100%.
- The uniqueness flag of the data fields can be controlled in the UI:



- If no information is available about the uniqueness of a data field, duplicate values are only reported as a quality problem if the majority of the values in the data field are unique. The confidence of the reported problem will be proportional to the percentage of unique values in the data field.

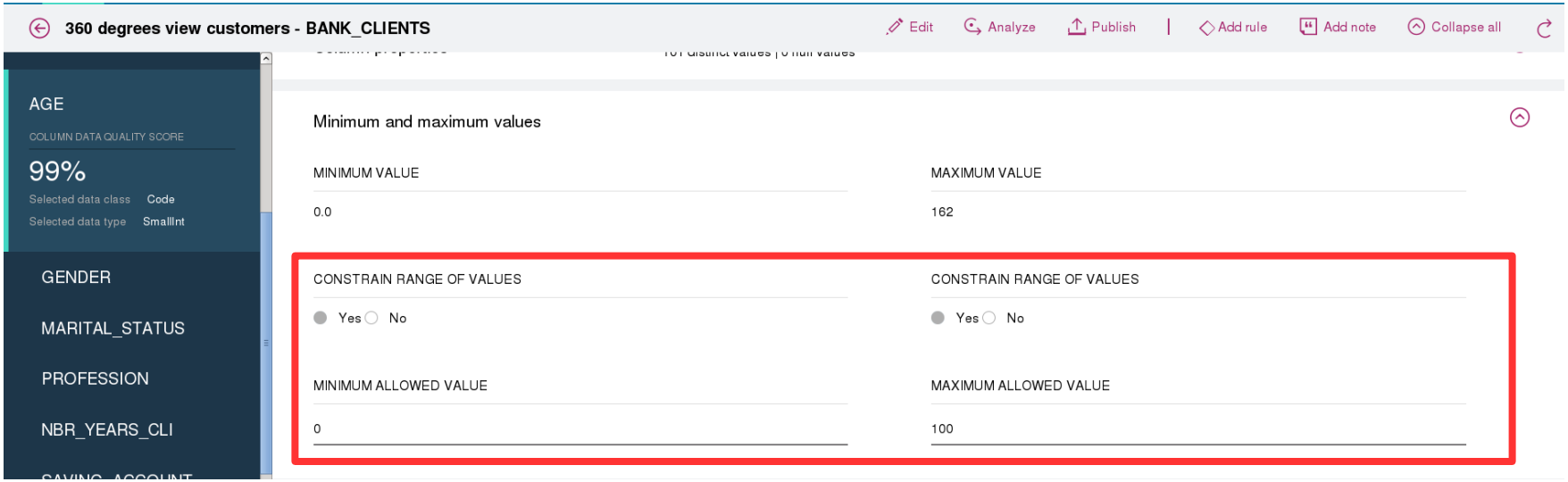
Data quality problems detected out of the box: Invalid formats

- Checks for values having a format marked by the user as invalid for its data field.
- The validity of the formats can be controlled in the UI:



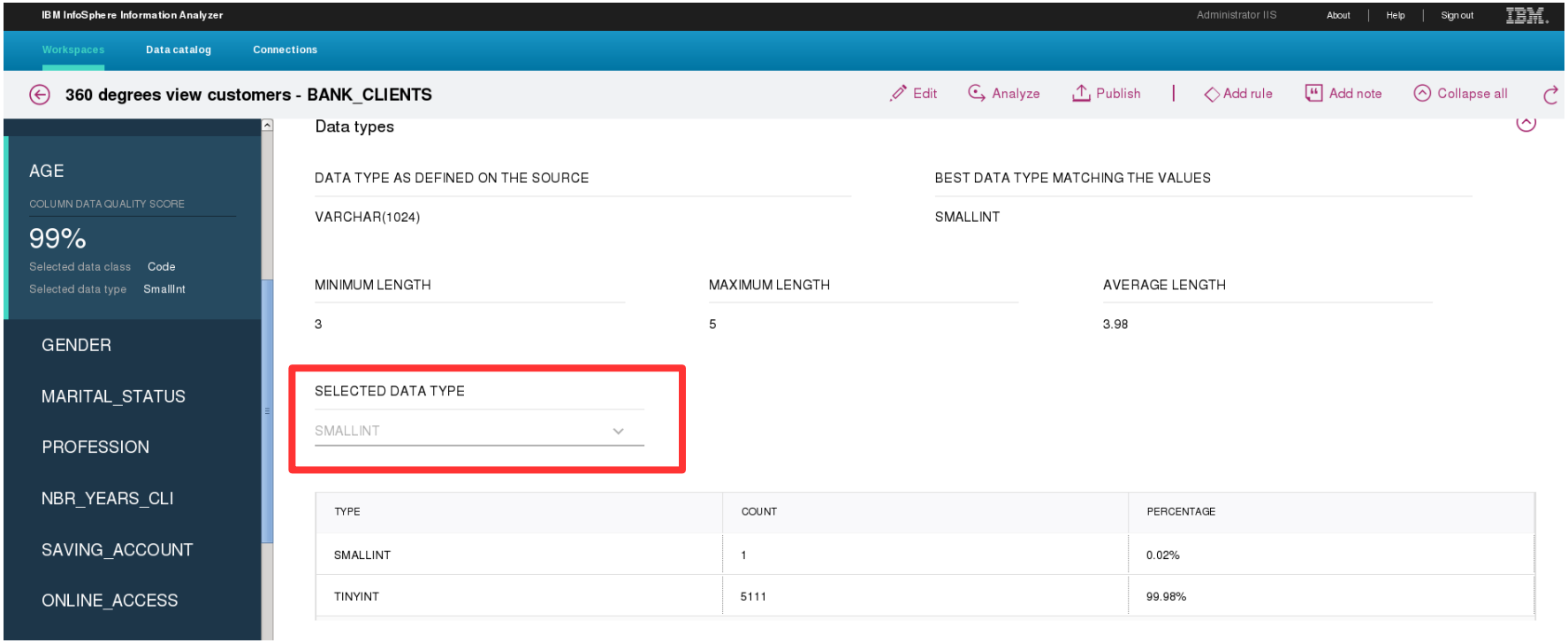
Data quality problems detected out of the box: Minimum / Maximum Range Violations

- Checks for values being outside the valid minimum/maximum range defined by the user for the data field.
- The validity of the formats can be controlled in the UI:



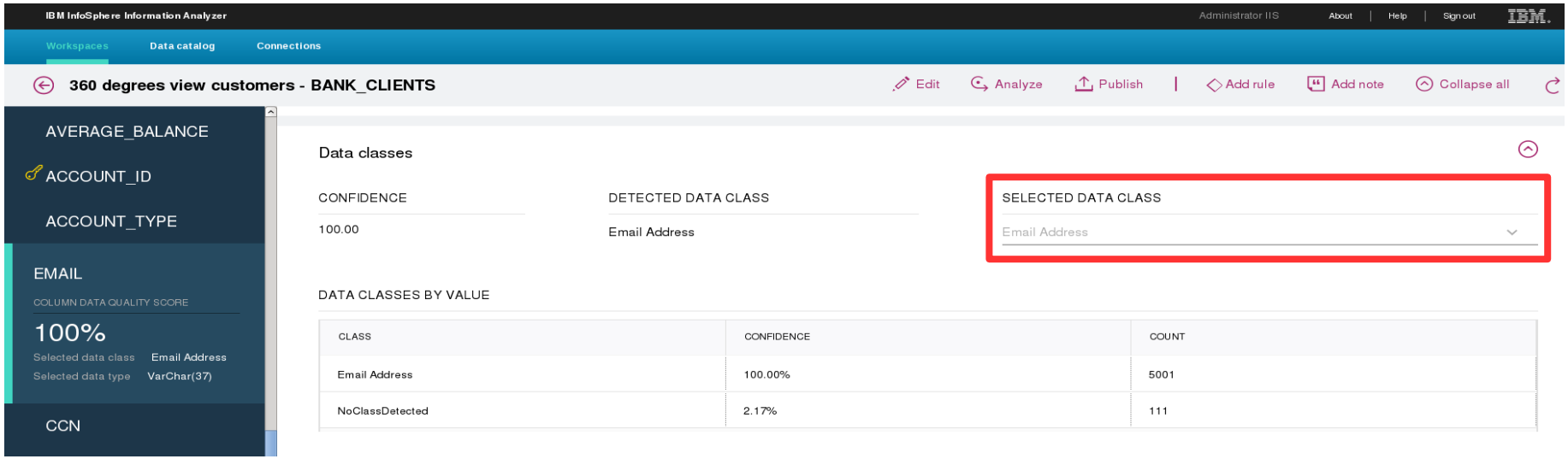
Data quality problems detected out of the box: Invalid Data Types

- Checks for values incompatible with the inferred/selected data type of their data field.
- The data type of the data fields can be specified in the UI:



Data quality problems detected out of the box: Data Class Violations

- Checks for values which don't match the data class selected for their data field.
- The selected data class of the data fields can be specified in the UI:



Data quality problems detected out of the box: Inconsistent use of upper and lower case

- Checks for values in a data field which have a different use of case (All upper case, vs. all lower case, vs. Name case, vs. Sentence case)
- Information Analyzer will recognize if the majority of the values of a data field use a consistent case usage, and mark values using a different case as a potential data quality problem.
- Ex:

Lucy V Adler
Cory J Gardner
Mary H Jacques
Shaun Q Dunda
Carol T Schwartz
HARRIS LAURENT

Data quality problems detected out of the box: Inconsistent representation of missing values

- Checks for data fields containing both NULL and empty of spaces values, as they indicate different and non standardized representations of the missing values.
- Ex:

NULL
NULL
NULL
<EMPTY>
<SPACES>
NULL
NULL
NULL

Data quality problems detected out of the box: Suspect Values

- Checks for values which are either numerical or categorical outliers, or values which seem not to be of the same domain as the other values of the same data field.
- For string values, the system will use advanced algorithms to detect patterns in the properties of the values, such as value length, % of letter/digits, nb of words, recurrent words, recurrent formats, used characters, etc... in order to detect values that look different than the majority of the values of the same data field
- Ex:

Name	Age
Lucy V Adler	32
Cory J Gardner	25
Mary H Jacques	18
Pdsaojfsadpoifj	46
Shaun Q Dunda	**156**
Carol T Schwartz	22

Data quality problems detected out of the box: Violations of Correlation

- Checks for values which are plausible in the context of their data field, but not in combination of the other values of the same row.
- Information Analyzer will identify numerical and categorical correlations between columns and mark the values that violate those correlations as potential data quality problems.
- Note: as this algorithm is more expensive in term of computation time, it is not enabled by default and needs to be enabled first:

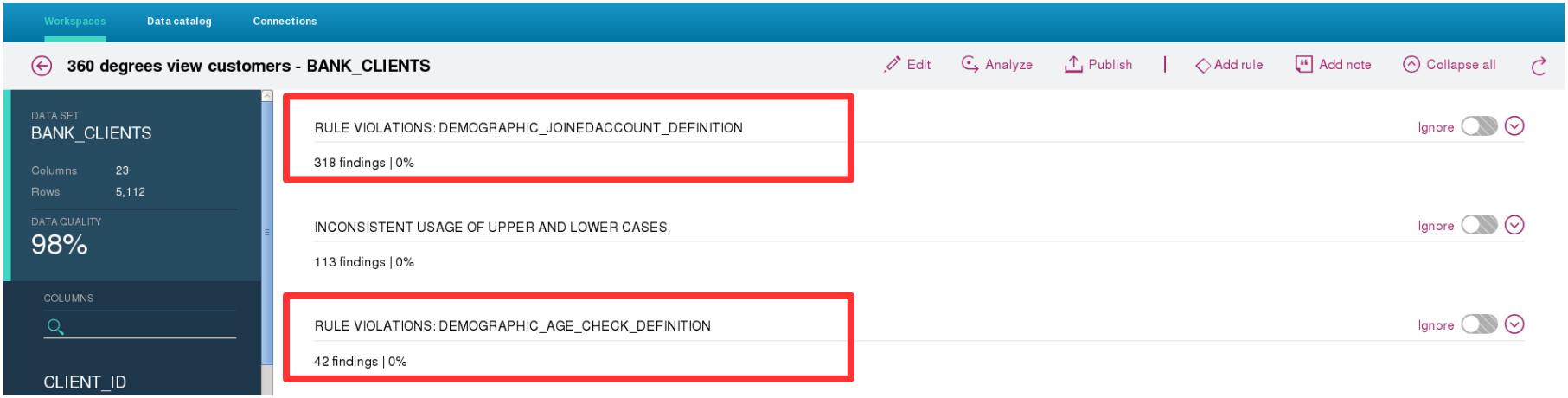
```
ASBServer\bin\iisAdmin.bat -set -key com.ibm.iis.ia.server.dqa.includeCorrelations -value true
```

- Ex:

City	State
San Francisco	CA
Las Vegas	NV
San Francisco	CA
San Francisco	**NV**
Las Vegas	NV
San Francisco	CA

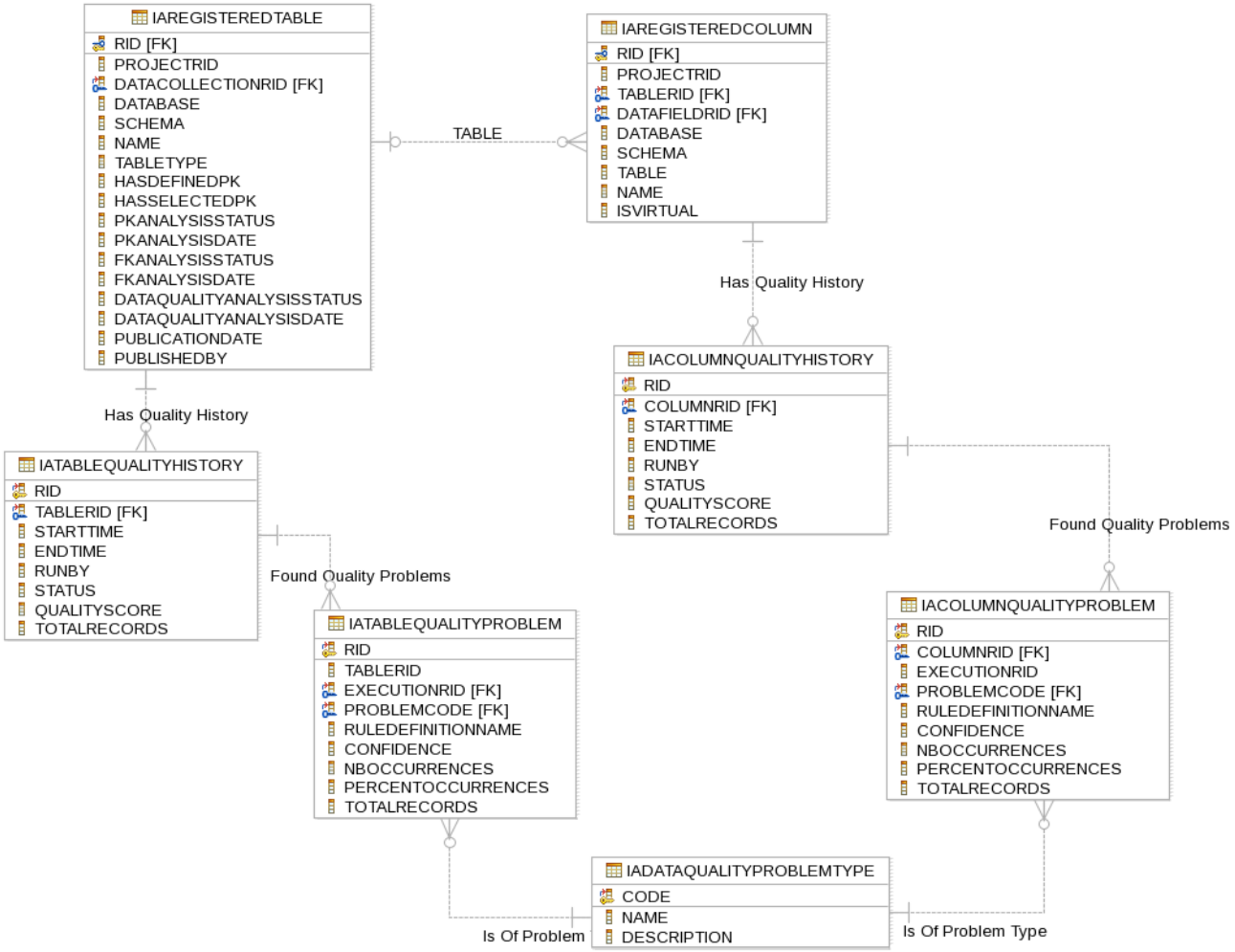
Data quality problems detected out of the box: Data Rule Violations

- If data rules or quality rules are defined on the analyzed data set, their logic is going to be automatically used during the data quality analysis if the rule fulfills the following conditions:
  - All its variables are only bound to data fields of the same data set or literals.
  - Its logic doesn't contain any lookup or aggregation.
- Each used data rule definition is reported as a separate data quality problem type.

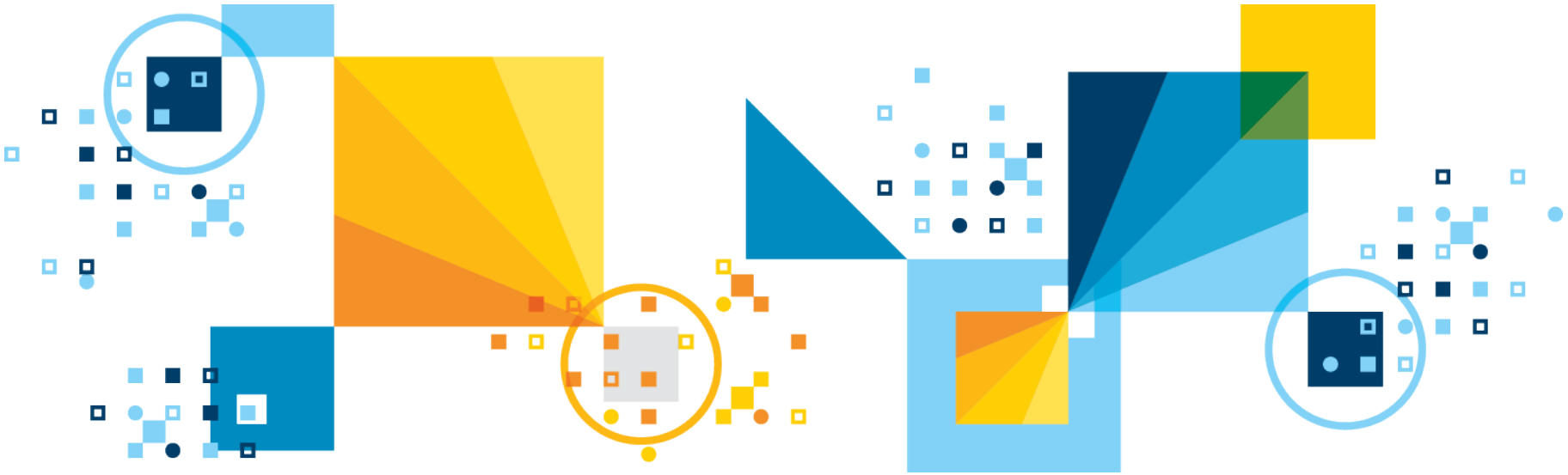


Reporting

- The computed data quality scores for each data set and data fields, as well as the details on the identified data quality problems can be queried by a set of new Xmeta SQL views.



Q&A



Thank You